# „Olcsó” morfológia

Novák Attila[1,2]

[1] MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport ,
[2] Pázmány Péter Katolikus Egyetem Információtechnológiai és Bionikai Kar,
1083 Budapest, Práter utca 50/a
{novak.attila}@itk.ppke.hu

**Kivonat** A számítógépes morfológiai leírások egy része a lexikon mellett szabálykomponenst is tartalmaz. Ez utóbbi biztosítja egyrészt a morfológiai leírás konzisztenciáját, másrészt megkönnyíti a morfológia új lexikai elemekkel való bővítését. Azonban egy ilyen típusú leírás elkészítése komoly erőfeszítést és különféle kompetenciákat igényel. A legtöbb szabadon elérhető morfológiai leírás viszont nem tartalmaz szabályokat. Ezek általában egy alaktani szótáron alapulnak, és a szavak lemmája és esetleg ettől eltérő töve mellett valamilyen a szó ragozási paradigmáját leíró információt tartalmaznak, gyakran valamiféle paradigmaazonosító címke formájában. Ezt esetleg még egyéb lexikai–szintaktikai–szemantikai információ egészítheti ki. Az ebben a cikkben bemutatott kutatás célja egy olyan algoritmus kidolgozása volt, amely lehetővé teszi, hogy a szabályalapú morfológiákhoz hasonlóan egyszerű módon lehessen az ilyen szótáralapú morfológiai leírásokba is új lexikai tételeket felvenni. A felügyelt tanításon alapuló algoritmus a szótárból hiányzó szavak helyes ragozási paradigmáját próbálja meg megjósolni a leghosszabb illeszkedő végződések és lexikai gyakorisági adatok felhasználásával. Az algoritmust orosz nyelvű adatokon mutatjuk be és értékeljük ki.

**Keywords:** morfológia, ragozási paradigma azonosítása, orosz

## 1. Bevezetés

A morfológiai elemzés a legtöbb természetesnyelv-feldolgozó rendszer fontos alapfeladata, amelyre sok más feldolgozási szint épül. Az információ-visszakeresési és szövegindexelési algoritmusok többsége is alkalmaz valamiféle morfológiai feldolgozást, mert a szöveget alkotó szavak lemmájának azonosítására szükség van a valóban használható kereséshez. Az utóbbi feladatok esetében ugyanakkor általában nincs szükség arra a morfoszintaktikai információra, amely az adott szóalak paradigmabeli helyét azonosítja, és amely a teljes körű morfológiai elemzés esetén a lemma és a szófaj mellett az elemzés részét képezi.

Az igényesen kidolgozott számítógépes morfológiákat általában olyan formalizmus használatával készítik el, amely a szavak morfológiai viselkedésének valamiféle szabályalapú leírását felhasználva minimalizálja az egyes lexikai tételekről a lexikonba felveendő információ mennyiségét. Ez egyrészt megkönnyíti az új lexikai tételek helyes felvételét a lexikonba, ugyanakkor lehetővé teszi, hogy a morfológia készítője teljesen ellenőrzése alatt tarthassa az általa létrehozott nyelvi

erőforrás minőségét. A szabályalapú morfológiai nyelvtanok létrehozása ugyanakkor többféle kompetenciát igényel: ismerni kell a formalizmust, az adott nyelv morfológiáját, helyesírását, morfofonológiáját, és kiterjedt lexikai ismeretekre van szükség. Sok számítógépes morfológiai adatbázis ugyanakkor nem tartalmaz külön szabálykomponenst. Ezeket az adatbázisokat általában valamilyen ragozási szótárban szereplő információ konverziójával hozzák létre. A szavak lemmája (és esetleg ettől eltérő töve) mellett valamilyen a szó ragozási paradigmáját leíró információt tartalmaznak (gyakran valamiféle paradigmaazonosító címke formájában), ezt esetleg még valamiféle egyéb lexikai–szintaktikai–szemantikai információval kiegészítve. Szabályok híján azonban az ilyen erőforrások új szavakkal való kiegészítése nem olyan egyszerű, mint a szabályalapú morfológiák bővítése. A gépi tanulás alkalmazása azonban lehetővé teheti, hogy a más morfológiákban a szabálykomponensben leírt tudást magából az adatbázisból kinyerve azt új szavak ragozási paradigmájának azonosításához használjuk. Módszerünk a tő különböző hosszúságú végződéseit és egyéb lexikai jellemzőit használja jellemzőkként a megfelelő ragozási paradigma kiválasztásához. Általában a leghosszabb illeszkedő végződésre leginkább jellemző morfológiai viselkedést veszi a legnagyobb súllyal figyelembe. Működését egy nyílt forráskódú orosz morfológiai lexikonon mutatjuk be és értékeljük ki.

Az automatikus paradigmaazonosítás lehetőségét a következő feladat megoldásával kapcsolatban vizsgáltuk meg. Egy szótárprogramot kellett képessé tennünk arra, hogy egy konkrét orosz-angol szótár szóanyagának összes ragozott alakját felismerje és helyesen lemmatizálja. A program a megszorításalapú morfológiai modellt használó Humor morfológiai elemzőt [1] használja a többi nyelv kezeléséhez, ezért az oroszhoz is ilyen elemzőt kellett készítenünk. Ahelyett azonban, hogy a semmiből hoztunk volna létre egy új orosz morfológiai adatbázist, a www.aot.ru címről letölthető LGPL-licenszű orosz morfológia [2] adaptálása mellett döntöttünk. Az erőforrást a Humor elemző által megkövetelt formátumra konvertáltuk. Ezután a szókincsét ki kellett egészítenünk a szótár azon a szavaival, amelyek az eredeti morfológiából hiányoztak.

Cikkünk felépítése a következő: a kapcsolódó munkák áttekintése után a 3. részben bemutatjuk a tanító- és tesztanyagként használt adatbázist. Ezt követően leírjuk azokat a jellemzőket, amelyeket az orosz esetében a szavak ragozási paradigmájának megjósolásához használtunk, majd részletesen bemutatjuk a tővégződéseket használó modellt és a paradigmajelöltek rangsorolását végző algoritmust. Végül a 7. részben kiértékeljük a rendszer teljesítményét, és áttekintjük a rendszer által elkövetett osztályozási hibák típusait.

## 2. Kapcsolódó munkák

A ragozási paradigmák automatikus azonosításával számos kutatás foglalkozott. Néhány tanulmány keretében a ragozási paradigmákat teljesen automatikusan nyers korpuszból próbálták megtanulni. A szóalakokat automatikusan csoportosították (clustering), és az így létrejött csoportokat elemezték [3,4,5]. További felügyelet nélküli morfológiatanuló rendszereket ír le Wicentowski [6], Hammar-

ström [7] és Goldsmith [8]. Az utóbbi munkában az azonos toldalékhalmazok szignatúráknak nevezett struktúrákba szerveződnek, amelyek ragozási paradigmákat reprezentálnak. A felügyelet nélküli módszerek teljesítménye mindazonáltal messze elmarad a felügyelt tanítást alkalmazóké mögött, dolgozzanak azok akár lexikai adatbázisok, akár elemzett korpuszok alapján.

Egy másik megközelítést alkalmazó kutatók ugyancsak használnak nyers korpuszokat. Ezekben a munkákban az adott szó számára megtippelt lehetséges ragozási paradigmák elemeit szóalak-gyakorisági adatokkal vetik össze az adott paradigmajelölt érvényességének ellenőrzéséhez [9,10]. Amennyiben az adott paradigma által jósolt szóalakok nem fordulnak elő a korpuszban, a paradigmajelöltet érvénytelenként elveti az algoritmus. Hasonló rendszert ír le Lindén [11], amely lexikai jellemzőkre és korpuszgyakorisági adatokra is támaszkodik a ragozási viselkedés analógiás meghatározásához. SVM-alapú osztályozót tanítanak be a Šnajder [12] által leírt alaki és gyakorisági jellemzőket egyaránt használó rendszerben annak a döntésnek a meghozatalára, hogy egy lehetséges paradigmajelölt elfogadható-e vagy sem.

Megközelítésünk a legtöbb itt leírt korábbi kutatástól különbözik abban, hogy kizárólag egy morfológiai lexikont használunk a ragozási paradigma automatikus megállapítását végző rendszerünk betanításához. Célunk a lemma és néhány egyszerű szótárakban is szereplő lexikai tulajdonság alapján a legvalószínűbb ragozási paradigma meghatározása. A szótárból származó információk alapján megbízhatóbban meg tudjuk jósolni a szótárba felveendő új szavak ragozási mintázatát, mint ha csak nyers korpuszadatok állnának rendelkezésünkre és mind a lemmát, mind az egyéb lexikai tulajdonságokat (pl. a szófajt), valamint a ragozási paradigmát is kizárólag ezek alapján kellene megjósolnunk.

## 3. A tanító- és a tesztanyag

Az itt leírt kísérletekhez az www.aot.ru [2] webhelyről letölthető LGPL-licenszű nyílt forráskódú morfológiai lexikont használtuk. A lexikon alapszókincse Zaliznyák ragozási szótárán alapul [13]. 174 785 lexikai tételt tartalmaz, amelyek mindegyike 2 767 ragozási paradigma valamelyikébe van besorolva. A paradigmaazonosító algoritmus kiértékeléséhez egy Serge Sharoff által készített orosz lemmagyakorisági adatbázist is használtunk.[3]

A morfológiai lexikont a lexikai tételeket a korpuszbeli lemmagyakoriságuk alapján csoportosítva háromféleképpen osztottuk tanító- és tesztanyagra. Az egyik csoportba a viszonylag ritka szavak kerültek. Ebbe a csoportba azok a lemmák kerültek, amelyek gyakorisága elérte a 8-at, de nem haladta meg a 10-et az internetes korpuszban (3970 szó). A második csoportba közepesen gyakori szavak kerültek, olyanok, amelyeknek lemmagyakorisága nem haladta meg a 100-at (36917 szó). A harmadik csoportba a nagyon gyakori szavak kerültek, 1000-nél nem kisebb gyakorisággal (9633 szó). Teszthalmazként ezeket a szóhalmazokat használtuk, tanítóhalmazként minden esetben az adott teszthalmaz teljes lexikonra vett komplementuma szolgált.

[3] http://corpus.leeds.ac.uk/frqc/internet-ru.num

## 4. Az orosz szavak paradigmatikus viselkedését befolyásoló tényezők

Amikor orosz szavak ragozási paradigmájának megjósolására teszünk kísérletet, bizonyos grammatikai jegyek ismeretére szükség van ahhoz, hogy helyesen tippelhessünk. A lemma és a szófaj egyértelműen ilyen jellemzők, bár a szófaj melléknevek és igék esetében általában igen jól megjósolható a lemma alakja alapján. Mindazonáltal ezeket a jellemzőket ismertnek tekintettük, hiszen minden szótárban szerepelnek.

A főnevek esetében emellett számos más lexikai/szemantikai jellemző is szerepet játszik annak meghatározásában, hogy milyen morfoszintaktikai jegyegyüttesek fordulnak elő egyáltalán az adott szó ragozási paradigmájában. Ilyen jegy a grammatikai nem, a megszámlálhatóság és az élőség. Emellett vannak ragozhatatlan főnevek. Ezek közül a jegyek közül a grammatikai nem minden szótárban szerepel. Emellett általában a ragozhatatlan főneveket is megjelölik. Bizonyos absztrakt, kollektív és anyagnév jelentésű szavaknak nincsenek többes számú alakjai. Másrészt csak többes számú alakokkal rendelkező főnevek is vannak. Az utóbbiak egy része az alakjuk alapján felismerhető: a lemmájuk tipikus a többes számú alakokra jellemző végződést visel.

Az élőség olyan formában befolyásolja a ragozási paradigmát, amely az adott főnév által felvett lehetséges alakok halmazára nincs hatással. A szó élő vagy élettelen mivoltától függően azonban a szó alakjai közül különbözőek esnek egybe. Az élők esetében a tárgyeset a birtokos esettel esik egybe (többes számban minden nemben, egyes számban csak hímnemben), a nem élők esetében a tárgyesetű alak az alanyesetűvel azonos. Ez a különbség élő–élettelen homonimapárok esetében is érvényesül. Ezt a jelenséget mutatjuk be a 1. ábrán a *ёж* sün: állat', és cseh sün: tankelhárító akadály' szavak ragozási paradigmájának összevetésével. Ugyanakkor, mivel az élőség jegy, bár az aot lexikonban szerepel, más szótárakban azonban általában nem jelölik,[4] ezért mi sem használtuk.

Hasonlóképp az igék esetében az, hogy a morfoszintaktikai jegyegyüttesek mely kombinációi érvényesek, az ige aspektusától és tranzitivitásától, illetve visszaható voltától függ. Például a nem tranzitív igéknek nincsenek passzív melléknévi igenévi alakjai, a befejezett aspektusú igéknek nincsenek jelen idejű melléknévi igenévi alakjai, és a folyamatos aspektusú igék legnagyobb részének nincsenek múlt idejű (különösképp passzív) melléknévi igenévi alakjai. Emellett az, hogy milyen határozói igenévi alakjai vannak egy igének, szintén az aspektuson, illetve egyéb idioszinkratikus lexikai tulajdonságokon múlik. Ezért ezeket a tulajdonságokat ismerni kell, és ezek az információk valóban szerepelnek a szótárakban.

A melléknév-ragozási paradigma defektivitásai, pl. a rövid predikatív alakok és a szintetikus közép- és felsőfokú alakok megléte különböző szemantikai és

[4] A szótár használójának a megadott jelentés alapján kell azonosítania ezt a tulajdonságot, ami általában sikerül is neki néhány szokatlan esetet kivéve: pl. a *мертвец* 'hulla' szó grammatikai szempontból élő (az ugyancsak 'hulla' jelentésű *труп* ugyanakkor élettelen).

| | |
|---|---|
| ёж[num:Sg.cas:Nom] | ёж[num:Sg.cas:Nom] |
| ежа[num:Sg.cas:Gen] | ежа[num:Sg.cas:Gen] |
| ежу[num:Sg.cas:Dat] | ежу[num:Sg.cas:Dat] |
| ежа[num:Sg.cas:Acc] | ёж[num:Sg.cas:Acc] |
| ежом[num:Sg.cas:Ins] | ежом[num:Sg.cas:Ins] |
| еже[num:Sg.cas:Prp] | еже[num:Sg.cas:Prp] |
| ежи[num:Pl.cas:Nom] | ежи[num:Pl.cas:Nom] |
| ежей[num:Pl.cas:Gen] | ежей[num:Pl.cas:Gen] |
| ежам[num:Pl.cas:Dat] | ежам[num:Pl.cas:Dat] |
| ежей[num:Pl.cas:Acc] | ежи[num:Pl.cas:Acc] |
| ежами[num:Pl.cas:Ins] | ежами[num:Pl.cas:Ins] |
| ежах[num:Pl.cas:Prp] | ежах[num:Pl.cas:Prp] |
| (a) ёж[N.gnd:Mas.ani:**Ani**][:8]; | (b) ёж[N.gnd:Mas.ani:**Ina**][:9]; |

1. ábra. A *ёж* sün' szó élő (a) és élettelen (b) jelentésű változatának ragozási paradigmája.

más látszólag idioszinkratikus jegyektől függenek. Például a relációs melléknéveknek általában nincsenek ilyen alakjai. Ezek a tulajdonságok ugyanakkor nem szerepelnek explicit módon az aot lexikonban, és általában a hagyományos szótárakban sem tüntetik fel őket, ezért a melléknevek esetében mi sem használtunk semmilyen lexikai jegyet a szófajon kívül.

Az adott lexikai tételhez ragozási paradigmát rendelő algoritmusunk számára tehát a lemma mellett a fent megadott lexikai tulajdonságokat (szófaj, nem, igeaspektus stb.) is hozzáférhetővé tettük. Ugyanakkor azok az információk, amelyek sem a hagyományos szótárakban nem szerepelnek, sem a szó alakjából nem jósolhatóak meg, nem szerepeltek az algoritmusunk számára ismert jellemzők között. Ilyenek voltak többek között, hogy egy főnév élő vagy élettelen dolgot jelent-e, hogy egy adott melléknévnek bizonyos alakjai léteznek-e, illetve hogy az adott szó ragozási paradigmájában bizonyos hangsúlyingadozások, helyesírási vagy egyéb rendhagyóságok szerepelnek-e.

A fent említett lexikai jegyek mellett algoritmusunk az adott lemma $n$ karakter hosszú végződéseit használta jellemzőkként különböző $n$ értékekre. A maximális $n$ végződéshossz paramétere az algoritmusnak. Kísérleteinkben ezt a paramétert 10-re állítottunk. A végződések és a lexikai jegyek által hordozott információ ábrázolásához toldalékmodellt építettünk a lexikon tanítóanyagként használt része alapján. Hogy ez a modell hogyan épül fel, azt a 2. ábra mutatja.

## 5. A végződésmodell létrehozása

A lemmavégződéseket és a lexikai tulajdonságokat a tanulóalgoritmus a 2. ábra jobb oszlopában bemutatott módon szófa adatszerkezetbe gyűjti. A lemmához az alábbi tulajdonságokat kódoló stringeket fűzi hozzá (jobbról balra haladva):

– A szögletes zárójelben álló címke két részből áll: a szófajból (és az alábbi példákban emellett a nemből), amelyet egy kötőjelet követően az adott szó ragozási paradigmájának az aot adatbázisban használt numerikus azonosítója követ. Ez az az információ, amelyet az algoritmusnak egy ismeretlen szóra meg kell tippelnie. A tanítóanyag feldolgozása során felépített szófa adatszerkezet végcsomópontjaiból kiinduló élek egy olyan adatszerkezetre mutatnak, ami az adott végződésű és lexikai tulajdonságokkal bíró szavakra a tanítóanyagban található [szófaj-paradigmaazonosító] párokból álló címkék eloszlását (relatív gyakoriságát) tartalmazza.
– A lemma végéhez attól egy függőleges vonallal elválasztva az adott lexikai elem ismert lexikai tulajdonságait kódoló stringet illesztünk.[5]
– Bizonyos paradigmákba tartozó lemmáknak egy adott végződést kell viselniük. Ilyen esetben a lemmában kettős kereszt jelöli annak a végződésnek a kezdetét, amely az adott paradigmaazonosító által jelölt paradigma alkalmazhatóságának a feltétele. Az adott paradigma biztosan nem jön szóba érvényes jelöltként olyan szavakra, amelyek nem az adott karaktersorra végződnek. Pl. minden a 1433-es számú paradigmához tartozó szó *ъё*-ra kell, hogy végződjön.

| | |
|---|---|
| `мумиё[N.n.*.-];prd:25` | `мумиё|n*[N.n-25]` |
| `остриё[N.n.-];sfx:ё;prd:1709` | `остри#ё|n[N.n-1709]` |
| `бабьё[N.n-];sfx:ё;prd:210` | `бабь#ё|ns[N.n-210]` |
| `дубьё[N.n-];sfx:ё;prd:210` | `дубь#ё|ns[N.n-210]` |
| `свежевьё[N.n-];sfx:ё;prd:210` | `свежевь#ё|ns[N.n-210]` |
| `цевьё[N.n.-];sfx:ьё;prd:1433` | `цев#ьё|n[N.n-1433]` |
| `жнивьё[N.n];sfx:ё;prd:1103` | `жнивь#ё|n[N.n-1103]` |
| `суровьё[N.n];sfx:ё;prd:210` | `суровь#ё|ns[N.n-210]` |
| `мостовьё[N.n];sfx:ё;prd:210` | `мостовь#ё|ns[N.n-210]` |

2. ábra. A végződésmodell egy részlete. A jobb oldali oszlopban álló elemek szerkezete: `lem#ma|lex-jegyek[Szófaj-ParadigmaID.]`, ahol a `ma` a lemma kötelező végződése minden olyan szó esetén, amely az adott numerikus `ParadigmaID` által azonosított paradigmába tartozik.

## 6. Rangsorolás

Az általunk használt toldalékszófa alapú rangsorolási algoritmust a Thorsten Brants TnT taggerében ([14]) a tanítóanyagban nem látott ismeretlen szavak lexikai valószínűségének becslésére használt toldalékguesser algoritmus ihlette. Azonban a Brants-féle algoritmus nem nyújtott kielégítő teljesítményt az általunk megoldani kívánt feladat esetében. Ezért addig módosítottuk az algoritmust, míg végül az eredetinél egyszerűbb, de annál lényegesen jobb eredménnyel

[5] n: semlegesnemű főnév, *: ragozhatatlan, s: csak egyes számú

működő modellt nem kaptunk. A paradigmabecslési algoritmus az adott szóhoz szóba jöhető paradigmák mindegyikéhez pontszámot rendel. Ez alapján a pontszám alapján rangsoroljuk a paradigma-jelölteket, és a legmagasabb pontszámút választjuk.

A pontszámot iteratív módon számítjuk ki az adott lemma ismert lexikai tulajdonságokkal bővített változatának végződésein végighaladva a legrövidebbtől a leghosszabb végződésig. Az iteratív pontszámszámítási algoritmus a 1. képlet szerint megadott módon módosítja az adott címkéhez tartozó pontszámot minden lépésben.

$$rank^{i+1}[tag] = sign \times len\_sfx \times rel\_freq + rank^{i}[tag] \tag{1}$$

ahol

| | |
|---|---|
| a $sign$ | negatív, ha a végződés rövidebb, mint az adott paradigma által megkövetelt minimális végződés |
| $len\_sfx$ | a végződés hossza a lexikai tulajdonságok nélkül |
| $rel\_freq$ | a $tag$ címke relatív gyakorisága az adott végződésre |
| $rank^{i}[tag]$ | -t elosztjuk $len\_sfx$-szal, ha $len\_sfx > 1$<br>negáljuk, ha $sign > 0$ és $rank^{i}[tag] < 0$<br>mielőtt a $rank^{i+1}[tag]$-t kiszámítanánk |

Ez a rangsorolási eljárás általában a leghosszabb illeszkedő végződéshez tartozó leggyakoribb paradigmát részesíti előnyben. A 3. ábrán néhány példát mutatunk be az algoritmus által rangsorolt paradigmajelöltekre.

```
рурба|f [N.f]    [N.f:50]#2.857270 [N.f:175]#0.756756 [N.f:48]#0.293840
                 [N.f:105]#0.175658 [N.f:88]#0.098045 [N.f:103]#0.051742
                 [N.f:396]#0.03995 [N.f:611]#0.039730 [N.f:69]#0.029693
                 [N.f:121]#0.021167

дурака|f [N.f]     [N.f:88]#4.466005 [N.f:15]#1.341181 [N.f:273]#0.904291
                   [N.f:36]#0.738748 [N.f:50]#0.467147  [N.f:16]#0.443249
                   [N.f:39]#0.300179 [N.f:105]#0.175658 [N.f:96]#0.155983
                   [N.f:103]#0.051742
```

3. ábra. A *рурба|f* és *дурака|f* inputra adott tíz legjobb paradigmajelölt. A jelölteket a pontszámuk alapján sorrendezi a rendszer. A pontszámot # jel választja el a javasolt címkétől.

## 7. Kiértékelés

A rangsoroló algoritmust a 3 részben leírt tesztanyagokon értékeltük ki. Ezeket a következőképpen jelöltük: ritka szavak (LT10), közepes gyakoriságú szavak (LT100), és gyakori szavak (MT1000). Módszerünk teljesítményének kiértékeléséhez a szokásos kiértékelési metrikákat alkalmaztuk.*A nyertes jelölt pontossága* azt adja meg, hogy az esetek mekkora részében rangsorolta az algoritmus a helyes paradigmát legelőre. Ez azt mutatja meg, hogy a rendszer mennyire jól rendel automatikusan paradigmaazonosítót egy adott szóhoz. Emellett a 2.–9. helyre rangsorolt azonosítók pontosságát is megmértük. A *fedés* azoknak a szavaknak az aránya, ahol a helyes paradigma benne van az első tíz helyre rangsorolt azonosítók halmazában. Lindén [11] megfontolásai alapján a hagyományos értelemben alkalmazott pontosság helyett a *maximális fedés melletti átlagos pontosság* mértékét használtuk. Ennek számítási módja $1/(1+n)$ minden szóra, ahol $n$ a helyes paradigma rangja a javaslatok között. Így a rangsoroló algoritmus minősége is mérhető. Mivel lehetséges olyan alkalmazás, ahol a paradigmaazonosítók automatikus meghatározása az emberi besorolást segítő alkalmazás csupán, ezért ez a metrika mutatja a zaj mértékét, amit az emberi validálás során ki kell szűrni. A fenti két metrika alapján meghatároztuk továbbá az *f-mértéket*, ami a pontosság és fedés harmonikus közepe.

Az algoritmus hatékonyságának meghatározásához két baseline rendszert állítottunk össze. Az első Brants toldalékguesser modelljét használja ([14]) a leghosszabban illeszkedő végződések helyett. Ebben a modellben szerepel egy $\theta$ tényező, amit a végződések alapján meghatározott címkevalószínűségek becslésének simításához használ. A $\theta$ tényező értékét a címkék valószínűségeloszlásának szórásával megegyező értékre állítja be. Először meghatározza az összes toldalék valószínűségi eloszlását a tanítóhalmaz alapján, majd a 2. képlet alapján szukcesszív approximációval simítja a modellt.

$$P(t|l_{n-i+1},...l_n) = \frac{\hat{P}(t|l_{n-i+1},...l_n) + \theta_i P(t|l_{n-i},...l_n)}{1+\theta_i} \tag{2}$$

minden $i = m...0$-ra, amelynek kezdeti értéke $P(t) = \hat{P}$, ahol

$\hat{P}$ a maximum likelihood becslések a lexikonbeli gyakoriság alapján

$\theta_i$ súlyok a címkék tanítóhalmazbeli feltétel nélküli maximum likelihood valószínűségének szórása minden $i$-re

A másik alkalmazott baseline rendszer minden szóhoz annak szófajcímkéje alapján a leggyakoribb paradigmaazonosítót rendeli hozzá. A két baseline rendszer teljesítményét hasonlítottuk össze a saját rendszerünkkel. Az összehasonítás eredménye a 1. táblázatban látható. Ahogy várható volt, a második baseline (amely egyszerűen a leggyakoribb paradigmát választja) nagyon alacsony pontosságot ért el. Saját módszerünk azonban az első baseline algoritmusnál is lényegesen jobban teljesít. Az utóbbi kettő közötti teljesítménykülönbséget az

magyarázza, hogy Brants modellje a feltétel nélküli, illetve a rövidebb végződések eloszlásának nagyobb súlyt ad, mint a hosszabbaknak. Ezzel szemben a hosszabb végződések alapján működő rendszerünk éppen fordítva működik, és így jobban figyelembe veszi az adott szóosztály viselkedését.

1. táblázat. A nyertes jelölt pontossága a paradigmaazonosítók esetén a leghosszabb toldalékillesztés módszere, Brants modellje és a leggyakoribb paradigma hozzárendelése esetén.

| | Leghosszabb toldalék | Brants modellje | Leggyakoribb paradigma |
|---|---|---|---|
| LT10 | 0.9166 | 0.6191 | 0.3464 |
| LT100 | 0.9039 | 0.6062 | 0.3403 |
| MT1000 | 0.7679 | 0.5372 | 0.3291 |

A paradigmákat és szintaktikai tulajdonságokat meghatározó címkék a szavak nagyon nagy felbontású osztályozását határozzák meg. Vannak olyan tulajdonságok, amik két paradigmát ugyan megkülönböztetnek egymástól, azonban a szavak ragozása szempontjából nem relevánsak. Például a nem ragozható határozószók számos alcsoportra oszlanak, de mindegyiknek csak egy alakja van. Ezek közül a tulajdonságok közül ráadásul a legtöbb nem is megjósolható. Sok esetben csupán a különböző hangsúlyingadozások tesznek különbséget két paradigma között, ami szintén nincs hatással a szavak leírt alakjára, azonban ilyenkor is más-más a helyes paradigmaazonosító. Végül, vannak olyan paradigmabeli különbségek is, amik az általunk megcélzott szótár-kiegészítési feladat szempontjából irrelevánsak, mert nincsenek hatással a paradigmában szereplő szóalakok halmazára. Ilyen például az azonos tövű élő és élettelen főnevek esete. Ezért, hogy a rendszerünk teljesítményét az eredetileg meghatározott tövesítési feladat szempontjából is kiértékelhessük, meghatároztuk a paradigmák ekvivalenciaosztályait. Ebben az esetben az automatikusan meghatározott paradigmaazonosítót helyesnek tekintettük, ha meghatározott paradigma által generált szóalakok halmaza megegyezett a helyes paradigma által meghatározott szóalakok halmazával. A 2 767 különböző közül a 921 nem egyedi paradigma 283 ekvivalenciaosztályba volt összevonható. A 2. táblázatban láthatóak az így mérhető eredmények, ahol a teljes' és 'equip' oszlopok a teljes paradigmaazonosító-egyezést megkövetelve, illetve az azonos ekvivalenciaosztályba tartozó paradigmák meg nem különböztetésével kapott eredmények. Megjegyzendő, hogy az 'equip' oszlopokban szereplő értékek összege nem 1, hiszen számos olyan eset fordul elő, ahol kettő vagy több olyan paradigma is szerepel az első helyekre soroltak között, amelyek a helyes paradigmával azonos szóalakhalmazt generál az adott lexikai tételhez.

Ahogy a számokból is látszik, a rendszerünk a ritka szavak esetében (LT10) teljesít legjobban, míg a gyakori szavak esetén mérhettük a legalacsonyabb teljesítményt (MT1000). Ez nem meglepő, hiszen a rendhagyó szavak a gyakoribbak

2. táblázat. A teljes címkeegyezés és az ekvivalenciaosztályok alapján elért eredmények

| | LT10 | | LT100 | | MT1000 | |
|---|---|---|---|---|---|---|
| | teljes | equip | teljes | equip | teljes | equip |
| #1 | 0.8924 | **0.9274** | 0.8750 | **0.9174** | 0.7416 | **0.8087** |
| #2 | 0.0614 | 0.2322 | 0.0685 | 0.2278 | 0.0684 | 0.2371 |
| #3 | 0.0168 | 0.2090 | 0.0223 | 0.2201 | 0.0314 | 0.2435 |
| #4 | 0.0057 | 0.1518 | 0.0078 | 0.1452 | 0.0168 | 0.1900 |
| #5 | 0.0035 | 0.1692 | 0.0037 | 0.1723 | 0.0090 | 0.2165 |
| #6 | 0.0015 | 0.1884 | 0.0019 | 0.1683 | 0.0083 | 0.1697 |
| #7 | 0.0000 | 0.1871 | 0.0012 | 0.1836 | 0.0032 | 0.1562 |
| #8 | 0.0005 | 0.1400 | 0.0011 | 0.1496 | 0.0043 | 0.1418 |
| #9 | 0.0010 | 0.1095 | 0.0007 | 0.1573 | 0.0017 | 0.1078 |
| pontosság | 0.9329 | 0.9538 | 0.92195 | 0.9481 | 0.8067 | 0.8550 |
| fedés | 0.9841 | 0.9876 | 0.9832 | 0.9875 | 0.8872 | 0.9158 |
| f-mértéke | 0.9578 | **0.9704** | 0.9516 | 0.9674 | 0.8450 | 0.8843 |

között fordulnak leginkább elő, míg a ritka szavak viselkedése ritkán rendhagyó, tehát jobban megjósolható. Nem meglepő, hogy a tanítóanyagban nem szereplő névmások vagy rendhagyó igék paradigmájának helyes meghatározása nem sikerül olyan jól. Ezen túl, mivel a célunk egy meglévő morfológiai lexikon kiegészítése, az ilyen lexikonok pedig a leggyakoribb szavakat eleve tartalmazzák, ezért ebben a feladatban éppen a ritka szavak helyes besorolása a fontosabb cél.

Szintén látszik az eredményekből, hogy hasonló fedésértékek mellett a pontosság és a nyertes jelölt pontossága szignifikánsan magasabb lett, amikor az ekvivalenciaosztályokat összevontuk. Az algoritmus tehát jól használható olyan erőforrások kiegészítésére, amelyeket teljes morfológiai elemzést nem igénylő, például információkinyerési vagy szótári kereséssel kapcsolatos feladatok megoldására használunk.

A 3. táblázatban a nyertes jelöltek pontossága látható szófajonkénti bontásban: az összes szóra, főnevekre, igékre és melléknevekre. Ebben az esetben a teljes paradigmacímkének való megfelelés helyett csak a paradigmaazonosítót vettük figyelembe. Így például a `[N.n._nam:Org.--49]`, `[N.n.--49]` és `[N.n--49]` javaslatok azonosnak tekinthetők. Az igék és melléknevek pontos paradigmájának meghatározása nehezebbnek bizonyult, mint a főneveké. Ennek okát elsősorban a következő fejezetben részletesebben tárgyalt szemantikai tényezők és hangsúlybeli különbségek között kereshetjük.

## 8. Hibaelemzés

A leggyakoribb tévesztések okai a leghosszabb végződést alkalmazó algoritmusunk esetén a ritka szavakra a következők: a rendszer nem tudja helyesen megjósolni, hogy

3. táblázat. A nyertes jelöltek pontossága minden szóra, illetve főnevekre, igékre és melléknevekre.

| | ÖSSZES | FŐNÉV | IGE | MELLÉKNÉV |
|---|---|---|---|---|
| LT10 | 0.9166 | 0.9547 | 0.8158 | 0.8665 |
| LT100 | 0.9039 | 0.9489 | 0.8114 | 0.8381 |
| MT1000 | 0.7679 | 0.8594 | 0.6884 | 0.5991 |

- egy melléknévnek vannak-e szintetikus középfokú alakjai
- a *-ние* végű elvont főneveknek van-e *-нье* alakú alternatív alakjuk
- egy főnévnek van-e második birtokos alakja (amelyet a partitívuszi szerkezetekben használnak)
- bizonyos igeosztályokban a múlt idejű melléknévi igenevek hangsúlya hogyan alakul (ez *e*∼*ё* váltakozáshoz vezet ezekben az alakokban – ugyanakkor a hétköznapi helyesírási gyakorlat ezt általában nem jelöli, tehát valójában ez nem vezet hibához)
- egy melléknévnek vannak-e szintetikus felsőfokú alakjai
- bizonyos melléknevek rövid és középfokú alakjainak hangsúlya hogyan alakul (ez szintén *e*∼*ё* váltakozáshoz vezet ezekben az alakokban)
- egy nem ragozódó főnév értelmezhető-e többes számúként
- egy folyamatos igének vannak-e múlt idejű melléknévi igenévi alakjai
- vannak-e a paradigmában szabad hangsúlyingadozások
- egy melléknévnek vannak-e rövid predikatív alakjai

A hangsúlyozással kapcsolatos és a szemantikából fakadó alakváltozatok hiányával járó esetek kivételével az algoritmus ritkán javasol helytelen paradigmákat. Az ismeretlen szavak esetén emberek is hasonló hibákat követnének el, különösen akkor, ha a szó jelentését sem ismerik. Továbbá rendszerünk az eredeti aot lexikonban szereplő inkonzisztenciákat is kimutatott, amelyek komolyabb orosz nyelvtudás híján is egyértelműen felismerhető hibák. Például, míg az *Кубаньэнерго* energiacég neve olyan megjelöléssel szerepel, hogy nincs többes száma, addig a hasonló *Сахалинэнерго* szónak nincs ez a tulajdonsága.

Az algoritmus által a gyakori szavak esetén vétett hibákat vizsgálva azok típusa hasonló. Mindazonáltal, ebben az adathalmazban a közép- és felsőfok, a második birtokos vagy második helyhatározói esetek helytelen meghatározása sokkal gyakoribb hibák, hiszen a gyakori szavak között sokkal nagyobb az ilyen „rendhagyó” alakok aránya.

Ugyanakkor a Brants-féle modellt használó algoritmus leggyakoribb hibái között olyan alapvető tévesztések szerepelnek, amiket kezdő orosz nyelvtanulók sem követnének el. Ez a rövid végződésekhez tartozó eloszlások túl nagy súllyal való figyelembevételéből ered, a hosszabbakkal szemben. Az első helyre rangsorolt paradigmajelölt gyakran egyáltalán nem alkalmazható az adott végződésű szavakra. A ritka szavakból álló tesztkorpuszon ennél a rendszernél a leggyakoribb

hiba például az, hogy *-ный* végű melléknevekre a *-кий* végűek paradigmáját javasolja alkalmazni.

## 9. Konklúzió

Jelen cikkünkben bemutattunk és kiértékeltünk egy toldalékszófa alapú felügyelt tanulási módszert alkalmazó algoritmust, ami ismeretlen szavak ragozási paradigmájának meghatározására alkalmas, azok lemmájának végződése és néhány lexikai tulajdonságuk alapján. A módszer a morfológiai szótárak alapján készített, és ezért szabálykomponenst nem tartalmazó számítógépes morfológiák lexikonának automatikus kiegészítésére használható. A módszer alkalmazhatóságát orosz nyelvre mutattuk be, azonban minimális adaptáció után az eszköz bármilyen más nyelvre alkalmazható, amihez rendelkezésre áll a megfelelő morfológiai erőforrás. Emellett éltünk azzal a feltételezéssel is, hogy a morfológia mellett létezik olyan elérhető szótár, amiben bizonyos lexikai tulajdonságok is megtalálhatóak, így ezeket a paradigmajavaslatok egyértelműsítése során figyelembe lehet venni. Módszerünk jól teljesít minden teszteset során, legjobban azonban a ritkán előforduló szavak esetén működik. Éppen ezek hiányoznak az eredeti lexikonból a legnagyobb valószínűséggel.

Az eredményekből az is kiviláglik, hogy a hosszabb szóvégződések előnyben részesítése a rövidebbekkel szemben lényegesen jobb teljesítményhez vezet. Ez akkor is világosan látszik, ha pusztán azt tekintjük, hogy milyen gyakran találja el az algoritmus pontosan a helyes paradigmát. Az elkövetett hibák elemzése azonban még inkább rávilágít a javasolt megoldás erősségeire. Míg a baseline toldalékguesser algoritmus az adott szóra egyáltalán nem alkalmazható paradigmákat javasol, mikor hibázik, addig az általunk bemutatott módszer csupán a szemantikai ismeret hiányából fakadó tévesztéseket követ el. Ezek azonban olyan hibák, amiket emberek ugyanígy elkövetnének.

## Hivatkozások

1. Novák, A.: What is good Humor like? [Milyen a jó Humor?]. In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
2. Sokirko, A.V.: Morphological modules at the site www.aot.ru. (2004)
3. Nakov, P., Bonev, Y., Angelova, G., Gius, E., von Hahn, W.: Guessing morphological classes of unknown German nouns. In Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R., eds.: RANLP. Volume 260 of Current Issues in Linguistic Theory (CILT)., John Benjamins, Amsterdam/Philadelphia (2003) 347–356
4. Monson, C., Carbonell, J.G., Lavie, A., Levin, L.S.: Paramor: Finding paradigms across morphology. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D., eds.: CLEF. Volume 5152 of Lecture Notes in Computer Science., Springer (2007) 900–907
5. Dreyer, M., Eisner, J.: Discovering morphological paradigms from plain text using a dirichlet process mixture model. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 616–627

6. Wicentowski, R.: Modeling and learning multilingual inflectional morphology in a minimally supervised framework. Technical report (2002)
7. Hammarström, H., Borin, L.: Unsupervised learning of morphology. Comput. Linguist. **37**(2) (2011) 309–350
8. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. Comput. Linguist. **27**(2) (2001) 153–198
9. Forsberg, M., Hammarström, H., Ranta, A.: Morphological lexicon extraction from raw text data. In: Proceedings of the 5th International Conference on Advances in Natural Language Processing. FinTAL'06, Berlin, Heidelberg, Springer-Verlag (2006) 488–499
10. Oliver, A., Tadic, M.: Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In: LREC, European Language Resources Association (2004)
11. Linden, K.: Entry generation by analogy – encoding new words for morphological lexicons. In: Journal Northern European Journal of Language Technology. (2009) 1–25
12. Šnajder, J.: Models for predicting the inflectional paradigm of Croatian words. In: Slovenščina 2.0. (2013) 1–34
13. Zaliznyak, A.A.: Russian grammatical dictionary – Inflection. Russkij Jazyk, Moskva (1980)
14. Brants, T.: TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, WA (2000)